Supplementary Tables

method	output transc. abund.?	bias correction options	details
mseq, Li et al.[1]	no	read start	Output is a bias model, which can be used to predict read starts on new transcript sequence. Trains a multiple additive regression tree (MART) on local sequence surrounding read starts, using those reads which align to single-isoform genes.
MISO, Katz et al.[2]	no	fragment length	Outputs and tests relative isoform abundance for multi-isoform genes. Bayesian model with Dirichlet prior on relative isoform abundance. Can incorporate fragment length distribution.
Cufflinks, Roberts et al.[3]	yes	fragment length, positional, read start	Fits a 21-bp variable length Markov model (VLMM) to the local sequence surrounding read starts using those fragments from single-isoform genes. Positional bias for 20 bins along transcript is fit for 5 transcript length classes.
RSEM, Li et al.[4]	yes	fragment length, positional	Fits a model of positional bias using the empirical distribution of read starts along transcripts.
Sailfish, Patro et al.[5]	yes	fragment length, read start	Previous versions offered post-hoc GC correction based on transcript GC content[6]. This has been deprecated and the latest version (0.9.0) offers read start bias correction to account for random hexamer priming. Same applies to *Salmon*.
kallisto, Bray et al.[7]	yes	fragment length, read start	Offers read start bias correction using a method based on that of *Cufflinks*.
Salmon, Patro et al.[8]	yes	fragment length, positional, read start	For *Salmon* 0.6.0, see note for *Sailfish*.
alpine	yes	fragment length, read start, positional, fragment GC, stretches of GC within fragment	Poisson generalized linear model to fit bias coefficients, see Methods.

Supplementary Table 1: Description of bias correction options offered by various methods under evaluation. See Figure 1a for diagram of various biases arising in RNA-seq. The second column reports whether the software provides transcript abundance estimates (one value for every transcript) as output. For methods with multiple citations, the citation above describes the bias correction methodology.

pop	center	assay	sample	experiment	run
TSI	UNIGE	NA20503.1.M_111124_5	ERS185497	ERX163094	ERR188297
TSI	UNIGE	NA20504.1.M_111124_7	ERS185242	ERX162972	ERR188088
TSI	UNIGE	NA20505.1.M_111124_6	ERS185048	ERX163009	ERR188329
TSI	UNIGE	NA20507.1.M_111124_7	ERS185412	ERX163158	ERR188288
TSI	UNIGE	NA20508.1.M_111124_2	ERS185362	ERX163159	ERR188021
TSI	UNIGE	NA20514.1.M_111124_4	ERS185217	ERX163062	ERR188356
TSI	UNIGE	NA20519.1.M_111124_5	ERS185167	ERX162948	ERR188145
TSI	UNIGE	NA20525.1.M_111124_1	ERS185212	ERX163022	ERR188347
TSI	UNIGE	NA20536.1.M_111124_1	ERS185156	ERX163042	ERR188382
TSI	UNIGE	NA20540.1.M_111124_2	ERS185349	ERX162940	ERR188436
TSI	UNIGE	NA20541.1.M_111124_4	ERS185125	ERX163043	ERR188052
TSI	UNIGE	NA20581.1.M_111124_4	ERS185181	ERX162937	ERR188402
TSI	UNIGE	NA20589.1.M_111124_3	ERS185057	ERX162793	ERR188343
TSI	UNIGE	NA20757.1.M_111124_1	ERS185169	ERX162732	ERR188295
TSI	UNIGE	NA20761.1.M_111124_7	ERS185420	ERX163049	ERR188479
TSI	CNAG_CRG	NA20524.2.M_111215_8	ERS185498	ERX162769	ERR188204
TSI	CNAG_CRG	NA20527.2.M_111215_7	ERS185082	ERX163033	ERR188317
TSI	CNAG_CRG	NA20529.2.M_111215_6	ERS185422	ERX162984	ERR188453
TSI	CNAG_CRG	NA20530.2.M_111215_6	ERS185442	ERX163025	ERR188258
TSI	CNAG_CRG	NA20534.2.M_111215_8	ERS185144	ERX162843	ERR188114
TSI	CNAG_CRG	NA20543.2.M_111215_5	ERS185134	ERX163170	ERR188334
TSI	CNAG_CRG	NA20586.2.M_111215_7	ERS185426	ERX162880	ERR188353
TSI	CNAG_CRG	NA20758.2.M_111215_8	ERS185342	ERX162819	ERR188276
TSI	CNAG_CRG	NA20765.2.M_111215_5	ERS185306	ERX162794	ERR188153
TSI	CNAG_CRG	NA20771.2.M_111215_7	ERS185108	ERX163165	ERR188345
TSI	CNAG_CRG	NA20786.2.M_111215_8	ERS185069	ERX162761	ERR188192
TSI	CNAG_CRG	NA20790.2.M_111215_6	ERS185378	ERX163152	ERR188155
TSI	CNAG_CRG	NA20797.2.M_111215_6	ERS185263	ERX162729	ERR188132
TSI	CNAG_CRG	NA20810.2.M_111215_7	ERS185427	ERX162968	ERR188408
TSI	CNAG_CRG	NA20814.2.M_111215_6	ERS185127	ERX163109	ERR188265

Supplementary Table 2: GEUVADIS samples used in this paper. CNAG_CRG was coded as center 1 and UNIGE was coded as center 2 in the text. The data consisted of paired-end reads of length 75 bp. Sequencing centers used the same library construction protocol for all samples. Full details of the library construction protocol can be found by searching the European Nucleotide Archive using any of the experiment IDs in the table (for example, searching for ERX163094)

Number of isoforms:	2	3	4	5-8	9-12	13+	Sum
Cufflinks changes	169	155	103	160	19	13	619
Cufflinks total	2867	1682	932	1096	142	42	6761
RSEM changes	173	157	107	163	25	12	637
RSEM total	2955	1687	947	1105	147	42	6883

Supplementary Table 3: Number of genes with changes in major isoform. Considering genes which have more than one isoform, and which had estimated FPKM greater than 0.1 in at least one isoform (total), shown is the number of genes for which the isoform with highest average FPKM was different across centers.

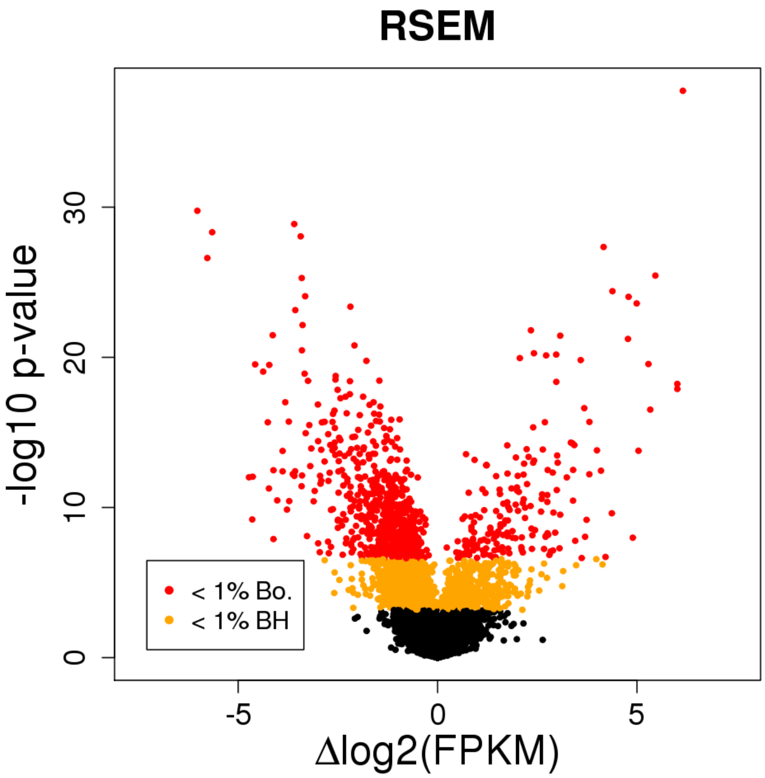
sample	site	library	runs
A	BGI	1	SRR:896663,896665,896667,896669,896671,896673,896675,896677
A	BGI	2	SRR:896679,896681,896683,896685,896687,896689,896691,896693
A	BGI	5	SRR:896727,896729,896731,896733,896735,896737,896739,896741
B	BGI	1	SRR:896743,896745,896747,896749,896751,896753,896755,896757
B	BGI	2	SRR:896759,896761,896763,896765,896767,896769,896771,896773
B	BGI	5	SRR:896807,896809,896811,896813,896815,896817,896819,896821
C	BGI	1	SRR:896823,896825,896827,896829,896831,896833,896835,896837
C	BGI	2	SRR:896839,896841,896843,896845,896847,896849,896851,896853
C	BGI	5	SRR:896887,896889,896891,896893,896895,896897,896899,896901
D	BGI	1	SRR:896903,896905,896907,896909,896911,896913,896915,896917
D	BGI	2	SRR:896919,896921,896923,896925,896927,896929,896931,896933
D	BGI	5	SRR:896967,896969,896971,896973,896975,896977,896979,896981
A	CNL	1	SRR:897047,897049,897051,897053,897055,897057,897058,897060
A	CNL	2	SRR:897062,897064,897066,897068,897070,897072,897073,897075
A	CNL	5	SRR:897107,897109,897111,897113,897115,897117,897118,897120
B	CNL	1	SRR:897122,897124,897126,897128,897130,897132,897133,897135
B	CNL	2	SRR:897137,897139,897141,897143,897145,897147,897148,897150
B	CNL	5	SRR:897182,897184,897186,897188,897190,897192,897193,897195
C	CNL	1	SRR:897197,897199,897201,897203,897205,897207,897208,897210
C	CNL	2	SRR:897212,897214,897216,897218,897220,897222,897223,897225
C	CNL	5	SRR:897257,897259,897261,897263,897265,897267,897268,897270
D	CNL	1	SRR:897272,897274,897276,897278,897280,897282,897283,897285
D	CNL	2	SRR:897287,897289,897291,897293,897295,897297,897298,897300
D	CNL	5	SRR:897332,897334,897336,897338,897340,897342,897343,897345
A	MAY	1	SRR:897407,897409,897411,897413,897415,897417,897419,897421
A	MAY	2	SRR:897423,897425,897427,897429,897431,897433,897435,897437
A	MAY	5	SRR:897471,897473,897475,897477,897479,897481,897483,897485
B	MAY	1	SRR:897487,897489,897491,897493,897495,897497,897499,897501
B	MAY	2	SRR:897503,897505,897507,897509,897511,897513,897515,897517
B	MAY	5	SRR:897551,897553,897555,897557,897559,897561,897563,897565
C	MAY	1	SRR:897567,897569,897571,897573,897575,897577,897579,897581
C	MAY	2	SRR:897583,897585,897587,897589,897591,897593,897595,897597
C	MAY	5	SRR:897631,897633,897635,897637,897639,897641,897643,897645
D	MAY	1	SRR:897647,897649,897651,897653,897655,897657,897659,897661
D	MAY	2	SRR:897663,897665,897667,897669,897671,897673,897675,897677
D	MAY	5	SRR:897711,897713,897715,897717,897719,897721,897723,897725

Supplementary Table 4: SEQC samples used in this paper. Sample A is Universal Human Reference RNA, sample B is Human Brain Reference RNA, sample C and D are a 3:1 mix and a 1:3 mix of A and B, respectively. Libraries 1 and 2 were prepared and sequenced at the site listed in the second column, while library 5 was prepared at a separate, fourth site and sequenced at the site listed in the second column. All the runs for a given experiment were combined to produce a single pair of FASTQ files. The data consisted of paired-end reads of length 100 bp.

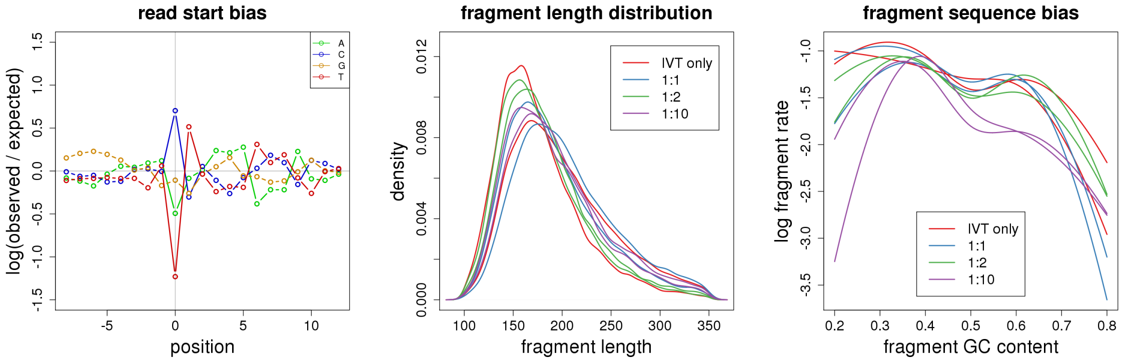
sample	protocol	experiment	runs
A	ribo-depletion	SRX307081	SRR:903050,903054,903056
A	ribo-depletion	SRX307082	SRR:903059,903065,903061
A	ribo-depletion	SRX307083	SRR:903066,903072,903069
A	ribo-depletion	SRX307084	SRR:903074,903077,903078
B	ribo-depletion	SRX307085	SRR:903087,903089,903088
B	ribo-depletion	SRX307086	SRR:903097,903090,903094
B	ribo-depletion	SRX307087	SRR:903101,903099,903103
B	ribo-depletion	SRX307088	SRR:903109,903106,903112
C	ribo-depletion	SRX307089	SRR:903114,903119,903120
C	ribo-depletion	SRX307090	SRR:903129,903122,903125
C	ribo-depletion	SRX307091	SRR:903137,903131,903133
C	ribo-depletion	SRX307092	SRR:903138,903140,903145
D	ribo-depletion	SRX307093	SRR:903147,903151,903153
D	ribo-depletion	SRX307094	SRR:903160,903155,903156
D	ribo-depletion	SRX307095	SRR:903165,903162,903167
D	ribo-depletion	SRX307096	SRR:903175,903177,903173
A	poly-A selection	SRX307097	SRR:903185,903178,903179
A	poly-A selection	SRX307098	SRR:903193,903192,903188
A	poly-A selection	SRX307099	SRR:903196,903201,903200
A	poly-A selection	SRX307100	SRR:903205,903206,903202
B	poly-A selection	SRX307101	SRR:903216,903211,903213
B	poly-A selection	SRX307102	SRR:903220,903223,903219
B	poly-A selection	SRX307103	SRR:903231,903228,903230
B	poly-A selection	SRX307104	SRR:903234,903238,903235
C	poly-A selection	SRX307105	SRR:903249,903247,903244
C	poly-A selection	SRX307106	SRR:903250,903254,903255
C	poly-A selection	SRX307107	SRR:903263,903259,903258
C	poly-A selection	SRX307108	SRR:903269,903268,903267
D	poly-A selection	SRX307109	SRR:903281,903277,903275
D	poly-A selection	SRX307110	SRR:903286,903288,903287
D	poly-A selection	SRX307111	SRR:903291,903294,903292
D	poly-A selection	SRX307112	SRR:903305,903304,903298

Supplementary Table 5: Information on the ABRF samples used in this paper. Sample A is Universal Human Reference RNA, sample B is Human Brain Reference RNA, sample C and are a 3:1 mix and a 1:3 mix of A and B, respectively. The runs for each experiment were combined to produce a single pair of FASTQ files. The data consisted of paired-end reads of length 50 bp.

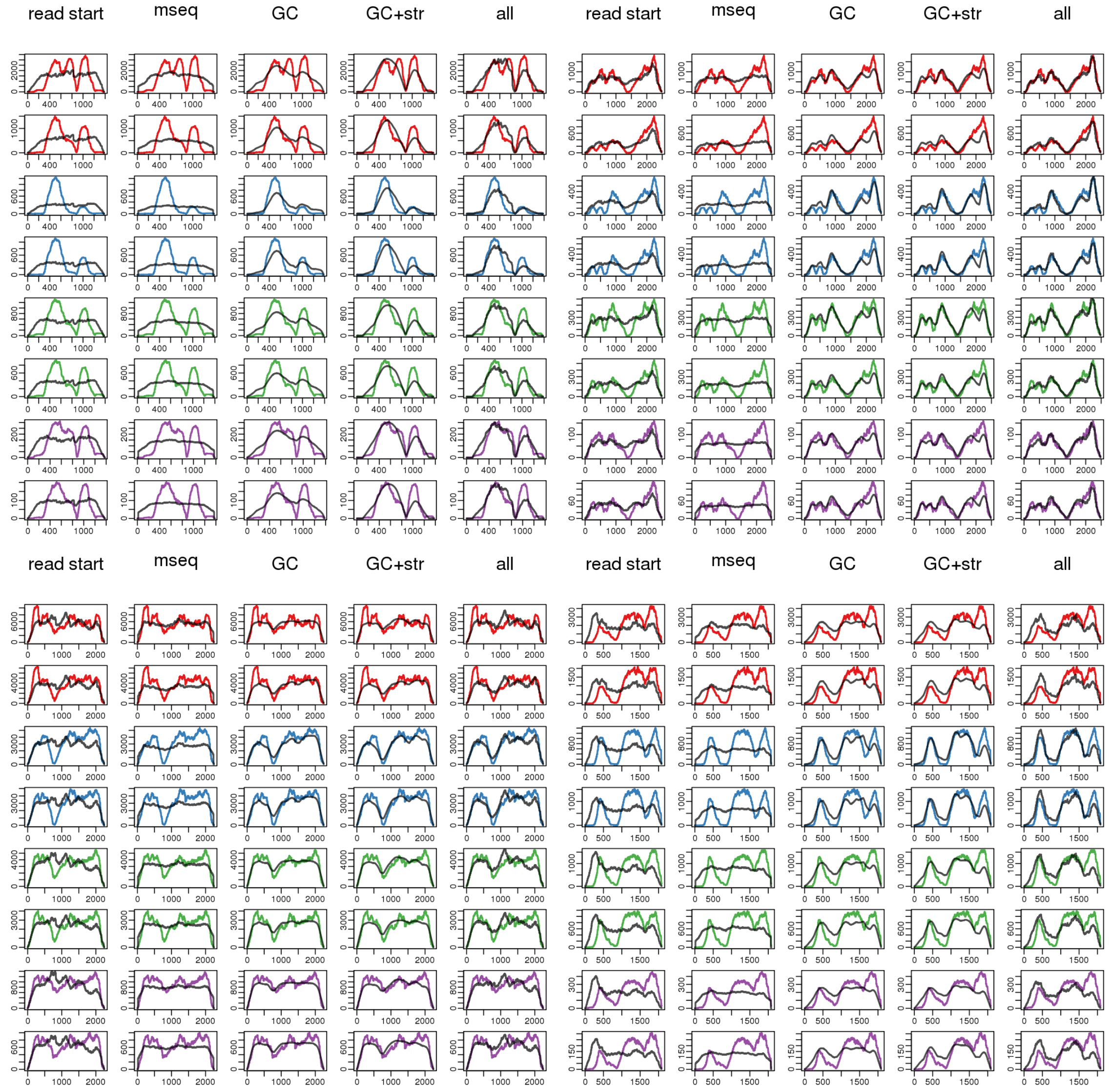
Supplementary Figures



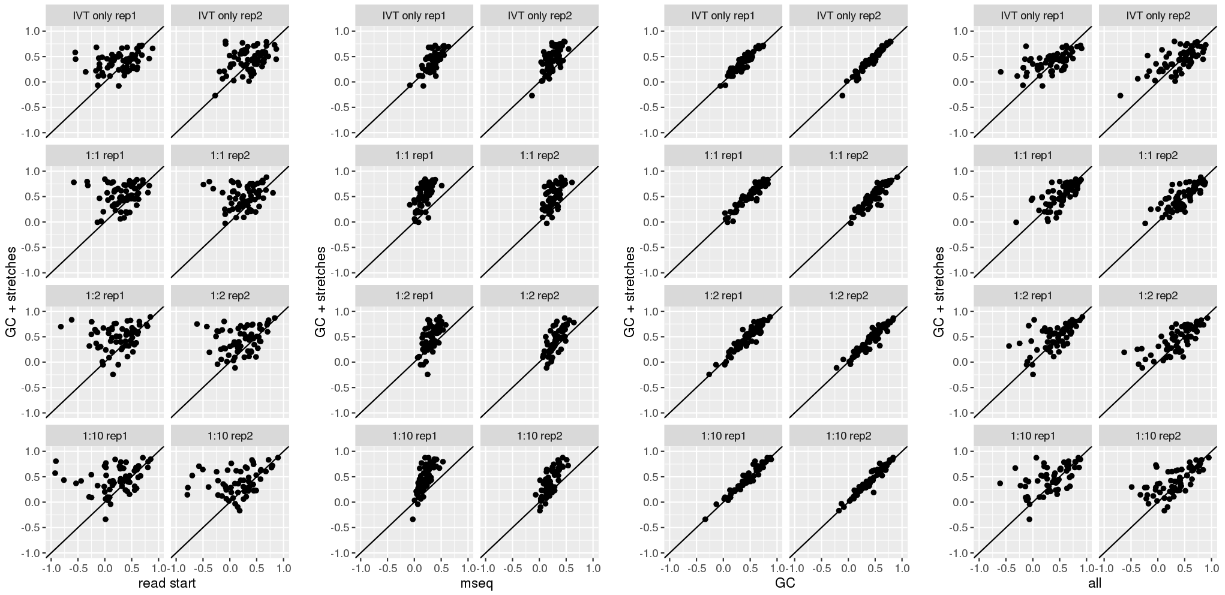
Supplementary Figure 1: Volcano plot of a comparison of *RSEM* transcript estimates across center. 2,829 transcripts had Benjamini-Hochberg adjusted p value less than 1% and 892 had family-wise error rate of 1% using a Bonferroni correction, out of 26,057 transcripts with FPKM estimate greater than 0.1.



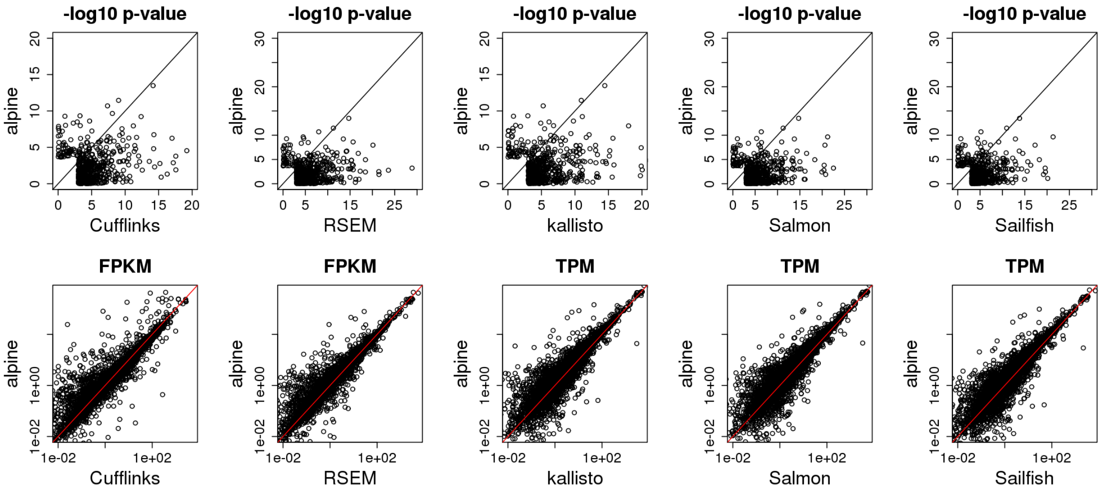
Supplementary Figure 2: Estimated bias parameters for the IVT-seq dataset. (Left) The 0-order terms of the read start bias model estimated for the 5' fragment end for one sample of the IVT-seq dataset. The 0-order terms are shown for visual simplicity, although the variable length Markov model (VLMM) used here and proposed by Roberts et al.[3] has higher order (1- and 2-order) Markov dependence for the middle positions. Both 5' and 3' end biases are combined for the read start bias calculation. (Middle) The estimated fragment length distributions for IVT-seq samples. (Right) The dependence of fragment rate on fragment GC content for IVT-seq samples. The curves were fit for the model with all terms, therefore representing the GC content dependence after removing read start bias and fragment length bias (which are pre-calculated and included as offsets, see Methods).



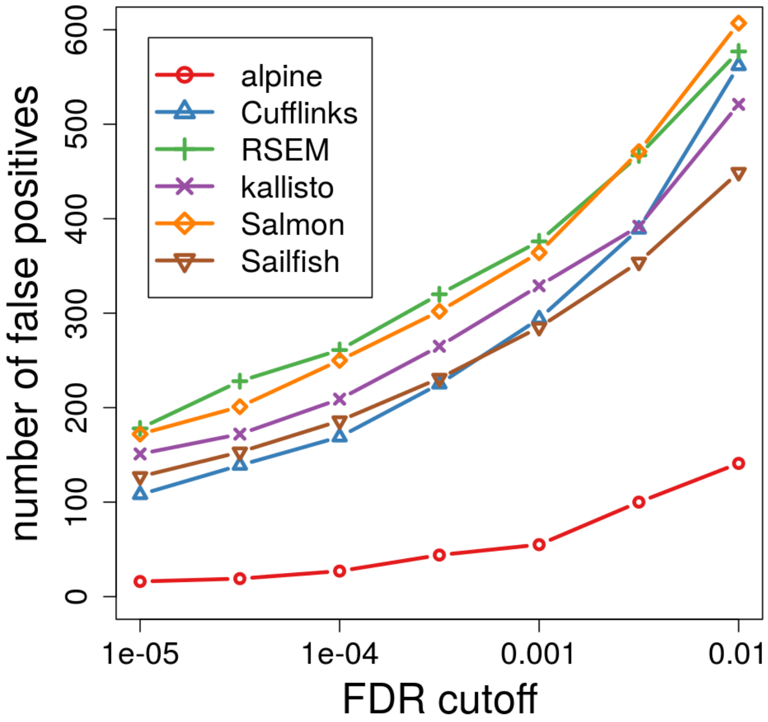
Supplementary Figure 3: Test set prediction of coverage for four IVT-seq transcripts for the following models: "read start": the *Cufflinks* VLMM for read starts implemented within *alpine*, the *mseq* model for read starts, "GC": a model using fragment GC content, "GC+str": as in "GC" plus additional terms for stretches of high GC within the fragment, "all": the VLMM for read starts in addition to the terms in "GC+str". Predicted coverage (black lines) and raw fragment coverage (colored lines) is shown for four different bias models and four transcripts: BC000158, BC011047, BC011377 and BC011380 (top left, top right, bottom left, bottom right), and for all eight samples (rows within each panel, where color denotes sample condition). Test set mean squared error was calculated by averaging the squared residuals from the predicted to the observed coverage.



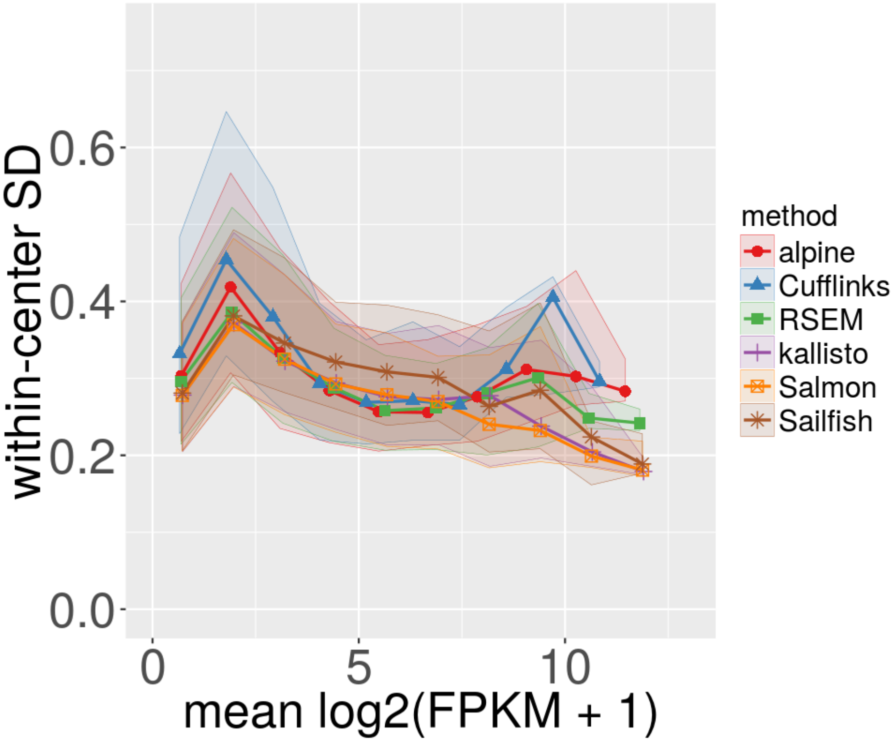
Supplementary Figure 4: Comparison of the reduction in test set mean squared error (MSE) for all 64 transcripts identified by Lahens et al.[9] for five models, split for each of eight IVT-seq samples. A value of 1 would indicate that all of the test-set error from a uniform coverage model was removed, while 0 would indicate the same error as a uniform coverage model. In all scatterplots, the y-axis shows the reduction in MSE for the model "GC + stretches" modeling both fragment GC content and high GC stretches. The x-axis shows (from left to right) the reduction in MSE for the *Cufflinks* VLMM for read starts, *mseq*x, the fragment GC content model (without GC stretches), and the model with the VLMM for read starts in addition to the terms in "GC + stretches". Points above the diagonal line indicate that the "GC + stretches" model outperforms the model on the x-axis.



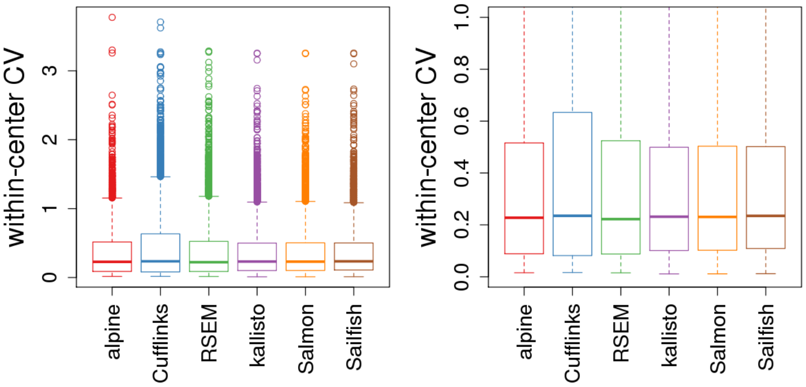
Supplementary Figure 5: Comparison of $-\log_{10}$ p values and expression estimates of *alpine* compared to other methods for the comparison across GEUVADIS sequencing center. p values and expression estimates are shown for the set of transcripts with adjusted p value less than 0.1 for either *alpine* or the method listed on the x-axis. *alpine* estimates of FPKM were converted into TPM for the rightmost three plots in the bottom row.



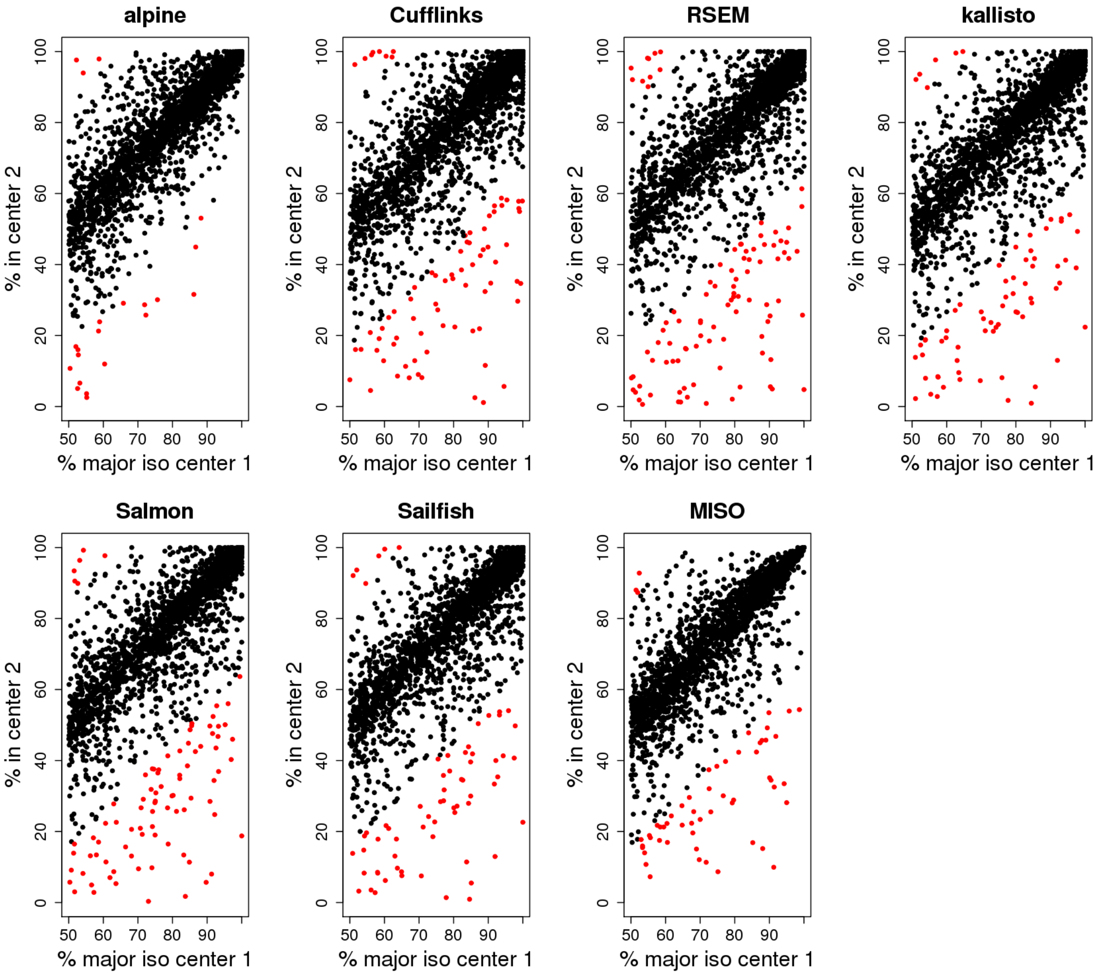
Supplementary Figure 6: Total number of false positives at various false discovery rate (FDR) cutoffs for the comparison across GEUVADIS sequencing center. Shown are the total number of transcripts reported as differentially expressed among 5,676 transcripts from genes with two isoforms, when comparing $\log_2(\text{FPKM} + 1)$ estimates of GEUVADIS samples across sequencing center. p values were adjusted using the Benjamini-Hochberg method. For *kallisto*, *Salmon*, and *Sailfish*, $\log_2(\text{TPM} + \text{PC})$ was used with a pseudocount corresponding to 1 on the FPKM scale.



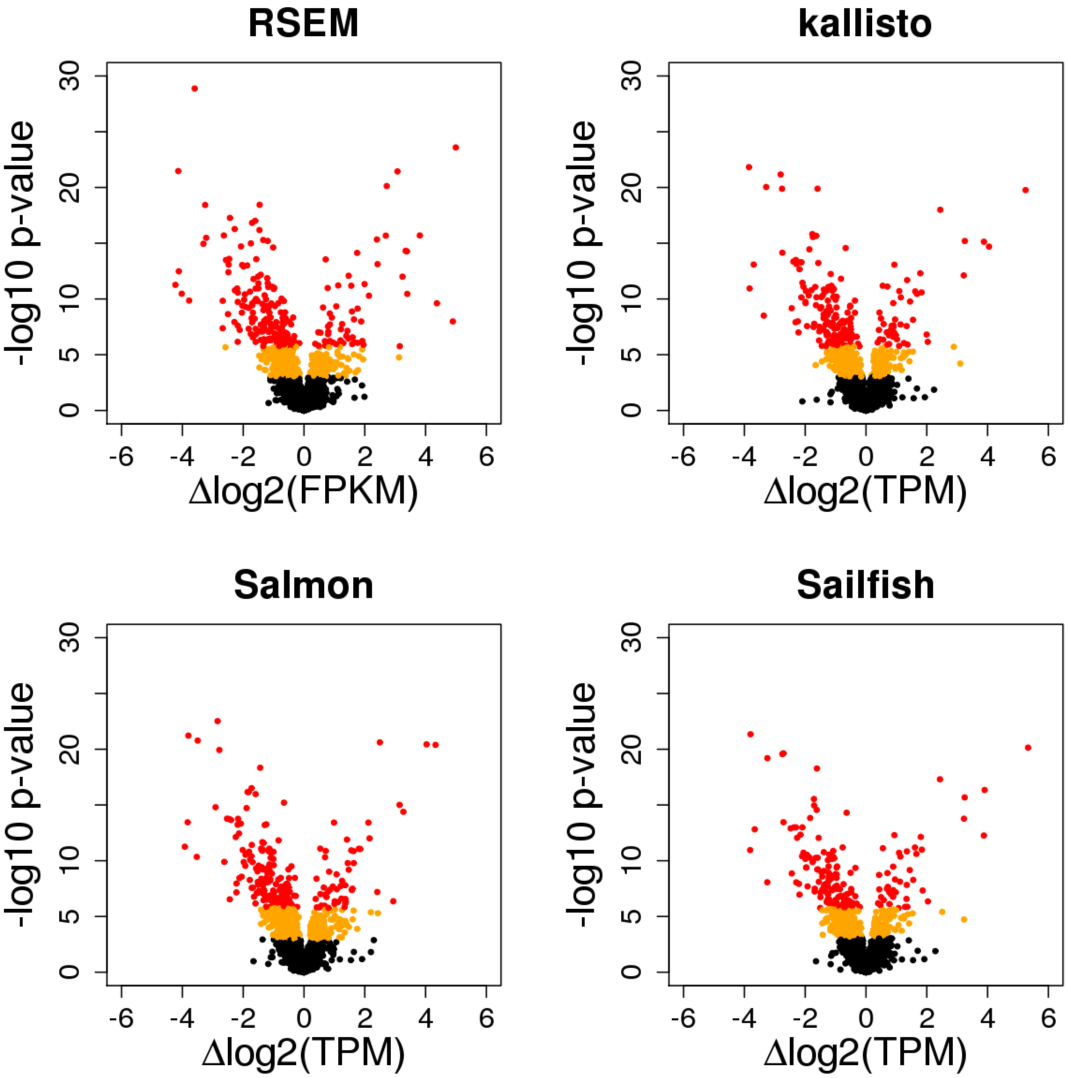
Supplementary Figure 7: Within-center standard deviation and mean of $\log_2(\text{FPKM}+1)$ estimates for the GEUVADIS dataset. Shown is the median (dark line), and 25% to 75% quantile (shaded region) of within-center standard deviation of transcript estimates for 10 bins along the mean. Only transcripts with FPKM estimates greater than 0.1 were included. For *kallisto*, *Salmon*, and *Sailfish*, TPM estimates were scaled to correspond to the FPKM estimates of *Cufflinks*.



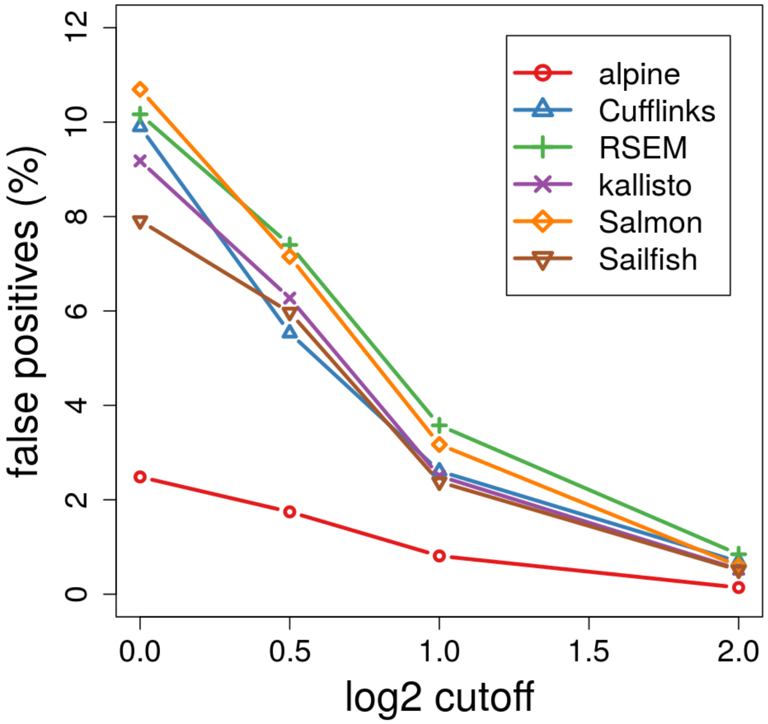
Supplementary Figure 8: Within-center coefficient of variation of $\log_2(\text{FPKM}+1)$ estimates for the GEUVADIS dataset. For transcripts with FPKM greater than 0.1, the within-center coefficient of variation (standard deviation divided by mean) was calculated for each center and averaged. TPM estimates were scaled to correspond to the FPKM estimates of *Cufflinks*. Right plot identical to left, with a cropped y-axis.



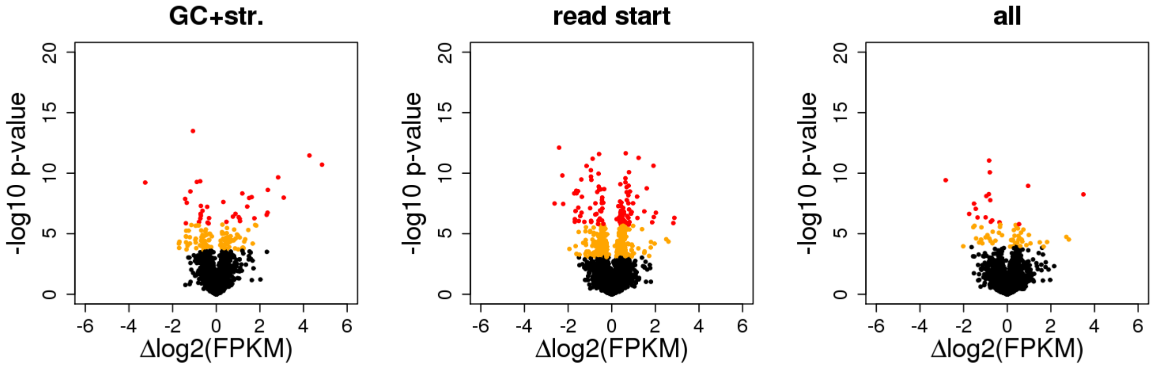
Supplementary Figure 9: Consistency of estimated percent expression of isoforms across GEUVADIS center for 2,838 genes with two isoforms in which one isoform had FPKM greater than 0.1 as reported by *Cufflinks*, and with at least 1 basepair overlap of isoforms. Estimated percent of the major isoform for center 1 shown on the x-axis (between 50% and 100% by definition). *alpine*, *Cufflinks*, *RSEM*, *kallisto*, *Salmon*, *Sailfish*, and *MISO* had 21, 76, 96, 70, 88, 66, and 56 genes with a change in estimated isoform percent greater than 35%, respectively.



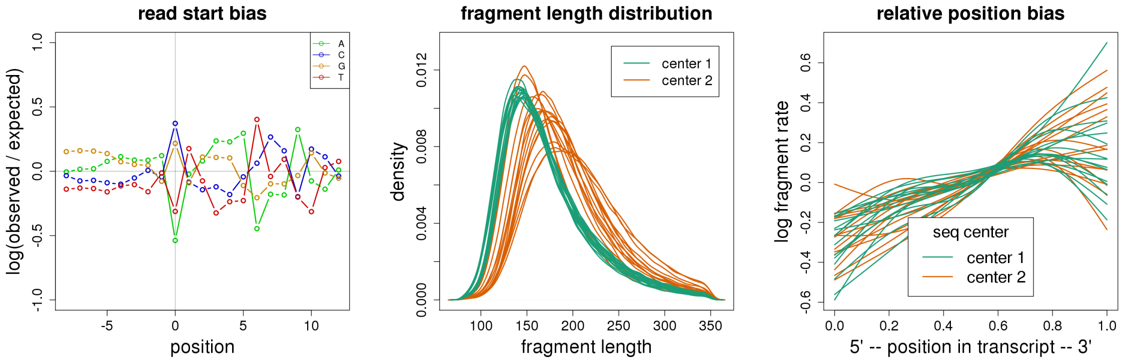
Supplementary Figure 10: Volcano plots of differential transcript expression across GEVUADIS center for genes with two isoforms, using *RSEM*, *kallisto*, *Salmon*, and *Sailfish* estimated FPKM or TPM. Out of 5,676 transcripts, these methods reported 577, 521, 607, and 449 transcripts with differential expression across center using an adjusted p value threshold of 1% (orange points), and 239, 187, 216, and 168 transcripts using a conservative Bonferroni family-wise error rate of 1% (red points), respectively.



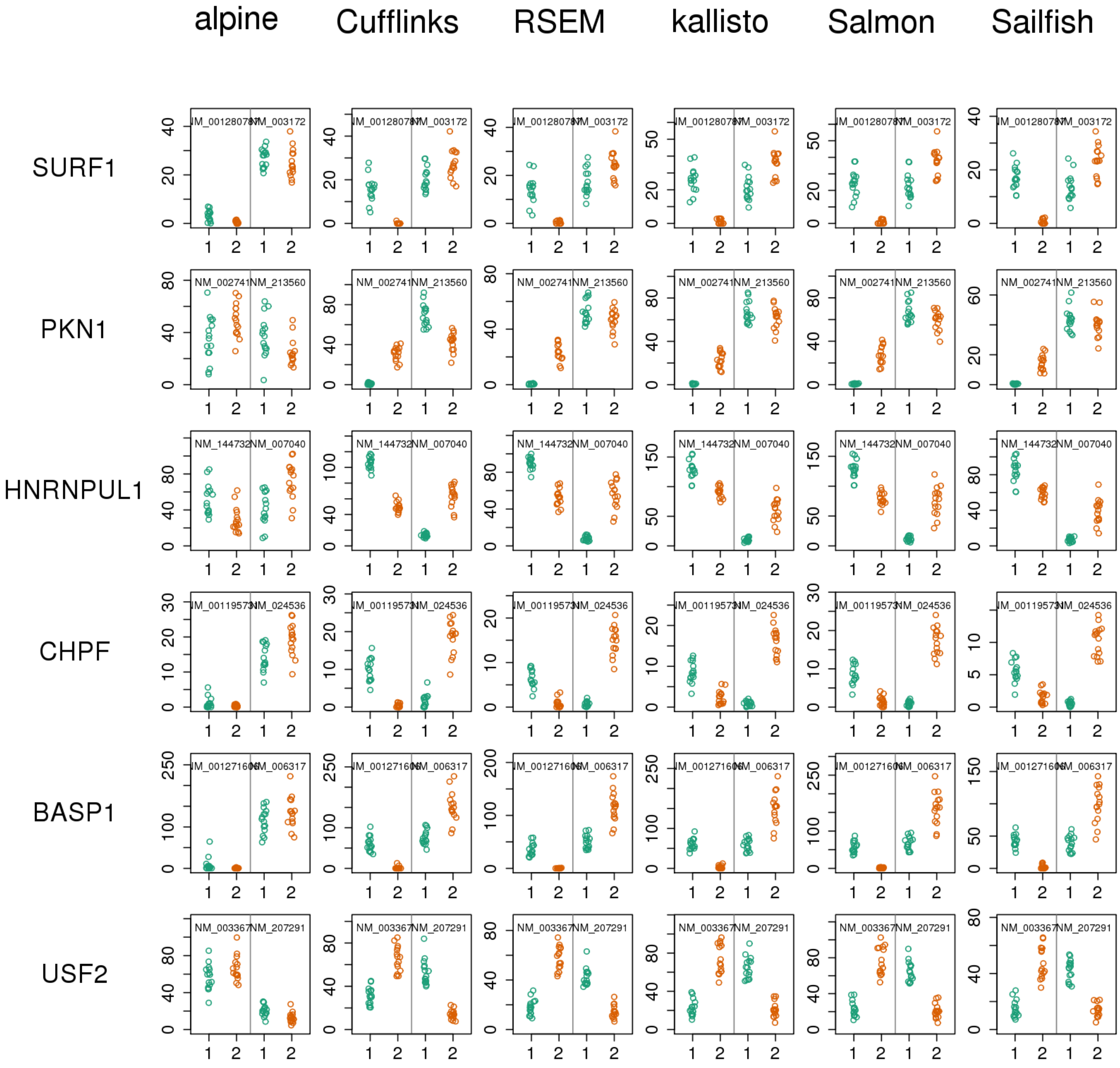
Supplementary Figure 11: Percent of false positives out of 5,676 transcripts at various $\log_2$ cutoffs for different methods. Here a false positive was defined as a transcript with Benjamini-Hochberg adjusted p value less than 1% and an estimated $\log_2$ fold change above a given threshold. Increasing the $\log_2$ cutoff reduced the false positives in the null comparison across sequencing center, but it should be noted that this procedure would also reduce sensitivity for a dataset containing true differences. In particular, adopting a $\log_2$ cutoff of 2, corresponding to a greater than fourfold change in transcript expression, would correspond to a great drop in sensitivity for many RNA-seq experiments.



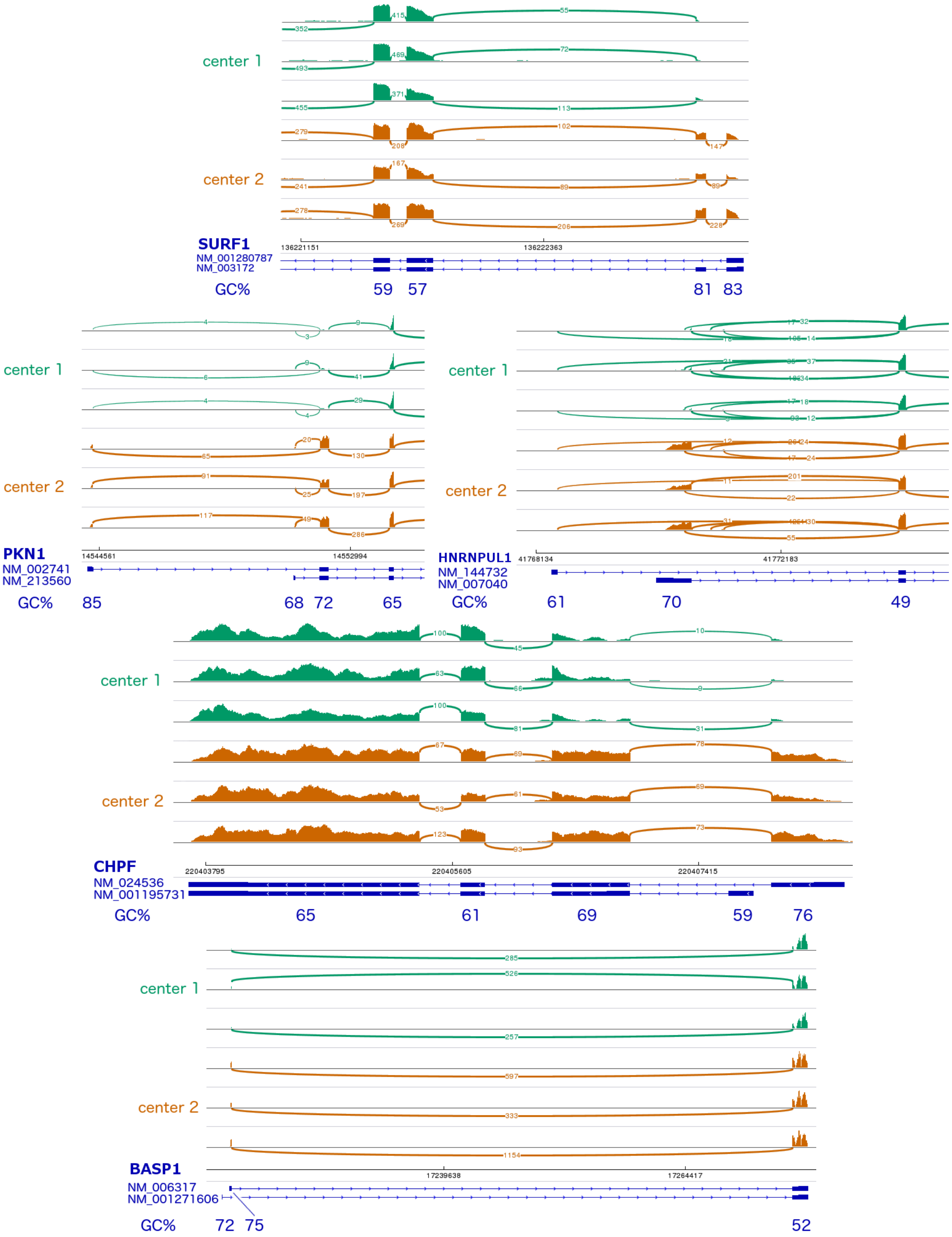
Supplementary Figure 12: Comparison of differences across GEUVADIS center for various *alpine* bias models. At a threshold on Benjamini-Hochberg adjusted p values of 1%, the models reported 141, 488, and 66 transcripts differentially expressed out of 5,676, for the models "GC+str." using fragment GC content and stretches of high GC within fragments, "read start" using the *Cufflinks* VLMM for read starts, and "all" using terms from the previous two models combined. The left-most plot is the same as shown in Figure 3d, repeated here for ease of comparison. At a more conservative Bonferroni threshold of 1% family-wise error rate, the models reported 37, 114, and 17 transcripts differentially expressed, respectively. See Methods for model details.



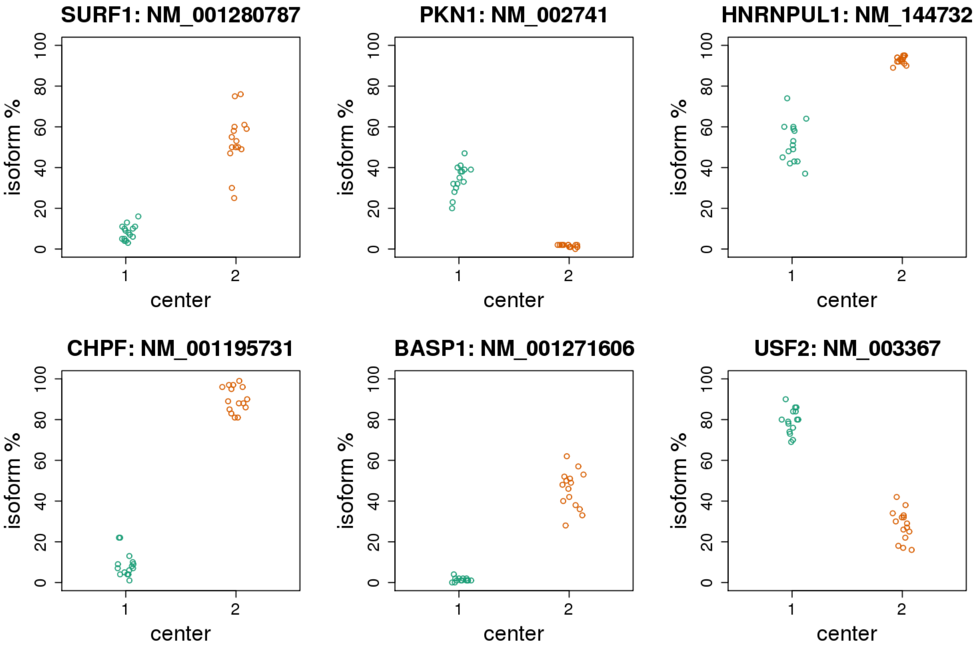
Supplementary Figure 13: Estimated bias parameters for the GEUVADIS dataset. (Left) The 0-order terms of the read start bias model estimated for the 5' fragment end for one sample of the GEUVADIS dataset. As in Supplementary Figure 2, the 0-order terms are shown for visual simplicity, although the variable length Markov model (VLMM) used here has higher order (1- and 2-order) Markov dependence for the middle positions. (Middle) The fragment length densities calculated for GEUVADIS samples. (Right) The relative position bias curves calculated for GEUVADIS samples. The fragment GC content curves for the GEUVADIS dataset are shown in the main text, in Figure 2e.



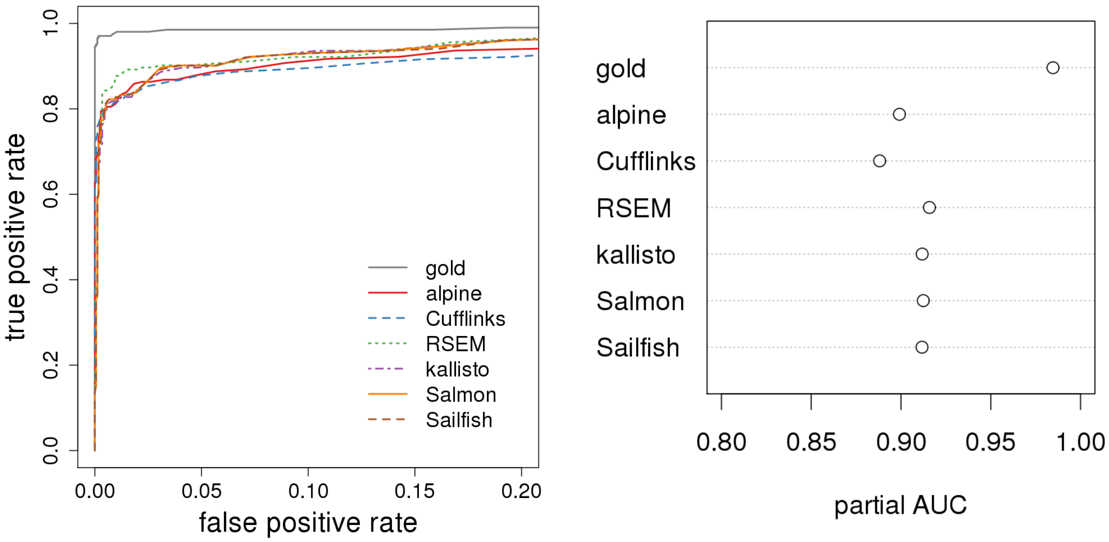
Supplementary Figure 14: Examples of estimated abundance across GEUVADIS sequencing center for genes with two isoforms, for six methods. Sequencing center (1 or 2) is indicated on the x-axis and estimated transcript abundance on the y-axis (FPKM for the first three methods and TPM for the last three methods). Examples selected for large across-center differences for *Cufflinks* and *RSEM*. Note that newer pseudoalignment-based methods *kallisto*, *Salmon*, and *Sailfish* had similarly discordant estimates of transcript abundance across sequencing center as the genome and transcriptome alignment-based methods *Cufflinks* and *RSEM*.



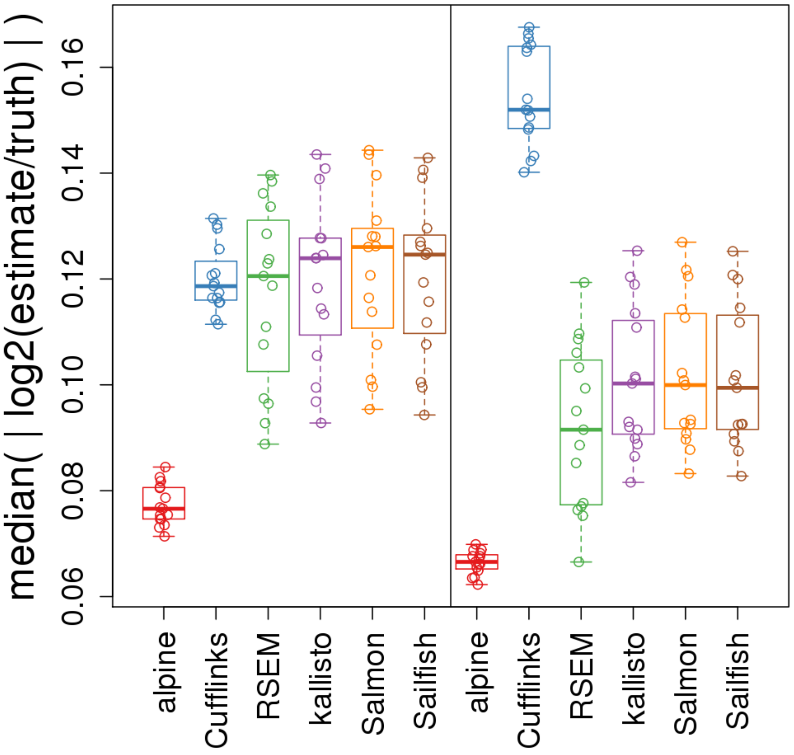
Supplementary Figure 15: Sashimi plots[10] for genes with two isoforms for which methods other than *alpine* predict isoform switching across sequencing center, as seen in Supplementary Figure 14. Sashimi plot for USF2 shown in Figure 2c. Regions of genes that distinguish isoforms are shown, except for CHPF and BASP1 where the entire gene is shown. In all cases, the abundance estimates of *alpine* were concordant with the qualitative evidence from junction-spanning reads: expression of a single isoform in SURF1, CHPF, and BASP1, and mixed isoform expression for PKN1 and HNRNPUL1.



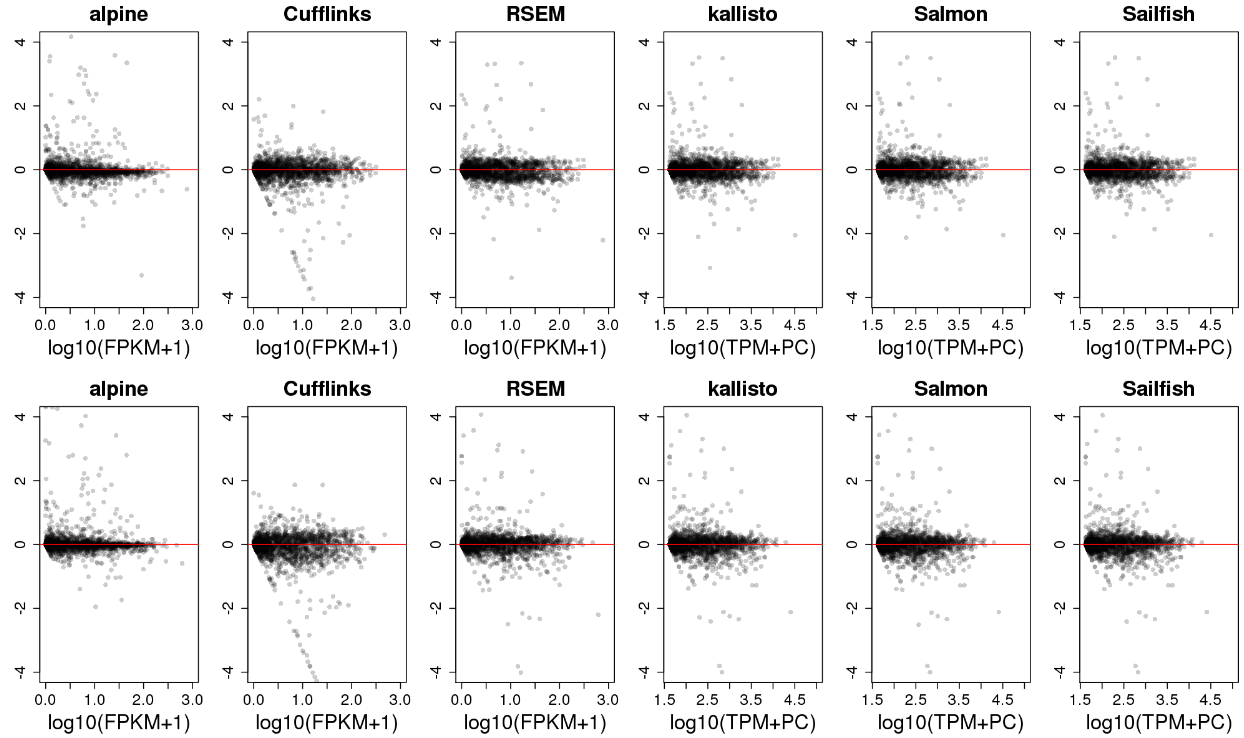
Supplementary Figure 16: *MISO* estimates of percent isoform expression were discordant across GEUVADIS sequencing center for the six examples genes with FPKM estimates shown in Supplementary Figure 14. *MISO* does not directly estimate transcript abundance, but provides estimation and testing on percent isoform expression. The y-axis shows the estimated relative abundance for one of the two isoforms out of total gene abundance. While the *MISO* method adjusts for varying fragment length distributions across samples, it was unable for these example genes to correct for the fragment GC bias that is modeled by *alpine*.



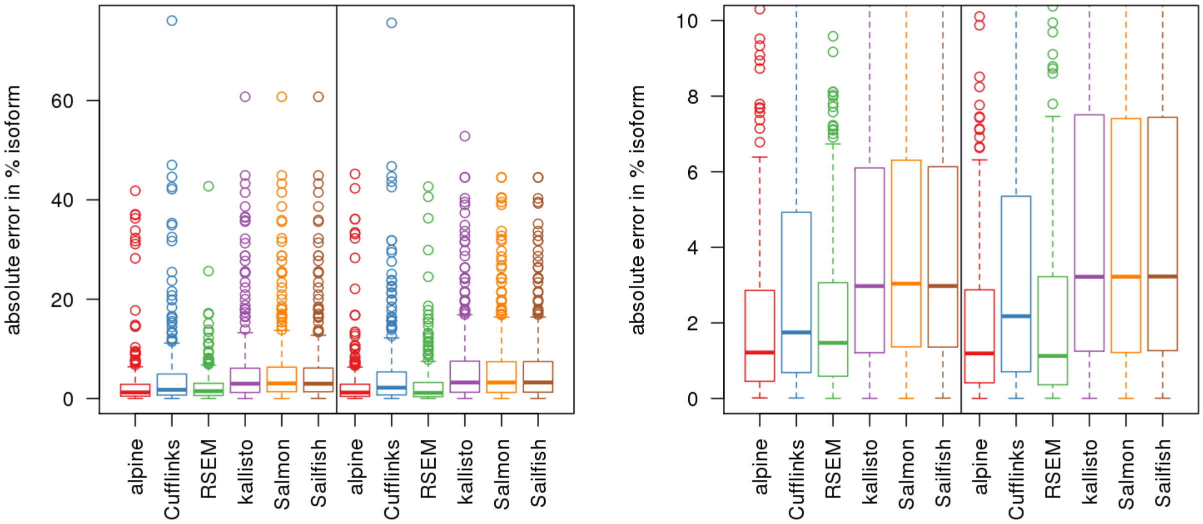
Supplementary Figure 17: ROC curve for the simulation of balanced design. A linear model was used to analyze expression values as in Figure 3g, except with the conditions balanced across sequencing center, and using a linear model with a blocking term indicating the sequencing center. Unlike in the confounded design, here experiment-wide sensitivity was not lost for the methods that do not model fragment sequence bias. Though the methods perform similarly here, with pAUC around 0.9, it should be noted that this is a best-case scenario for the competing methods, as the batches were known, the residual degrees of freedom were high, and therefore batch effects could be removed from transcript abundance estimates by adding a blocking term to the linear model. However, in the more common scenario in which batches are not known, this approach is not applicable. In contrast, our approach of removing sample-specific fragment sequence bias by estimating it directly produces (i) accurate within-sample estimates of transcript abundance and isoform percentages and (ii) reduces false positives for relative abundance across samples in the case of total or partial confounding, even when the batches are not known, or if sample-specific deviations exist within batches. The right panel displays the partial AUC for the ROC curves with false positive rate range in $[0, 0.2]$, The partial AUC was scaled to take values between 0 and 1.



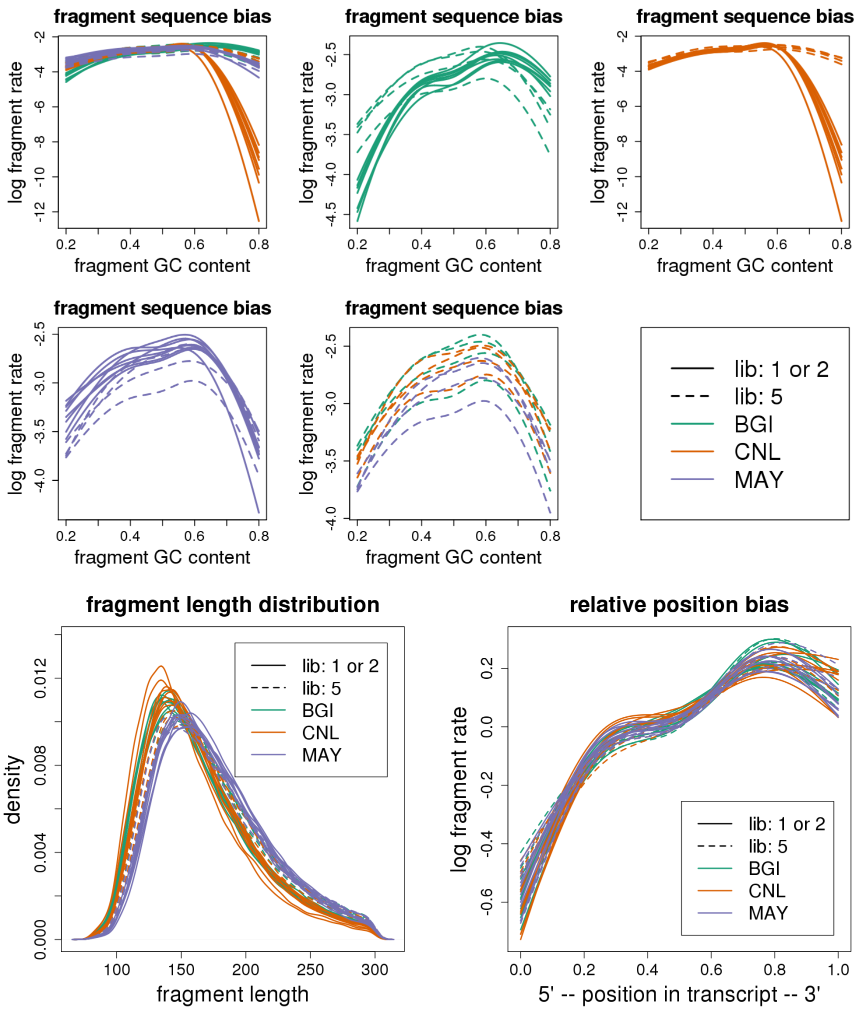
Supplementary Figure 18: Median absolute $\log_2$ fold error in estimating transcript expression levels for all simulated samples, split by sequencing center. The left side provides the median error rate over transcripts for those samples from GEUVADIS sequencing center 2 (orange curves in Figure 2e), which had less dependence of fragment rate on fragment GC content. The right side provides the median error rate over transcripts for samples from center 1 (green curves in Figure 2e), which had strong dependence of coverage on fragment GC content. *kallisto*, *Salmon*, and *Sailfish* TPM estimates were compared to gold-standard TPM values. *alpine* had the lowest median absolute error. For an example of $\log_2$ fold error over true expression values for individual simulated samples, see Supplementary Figure 19.



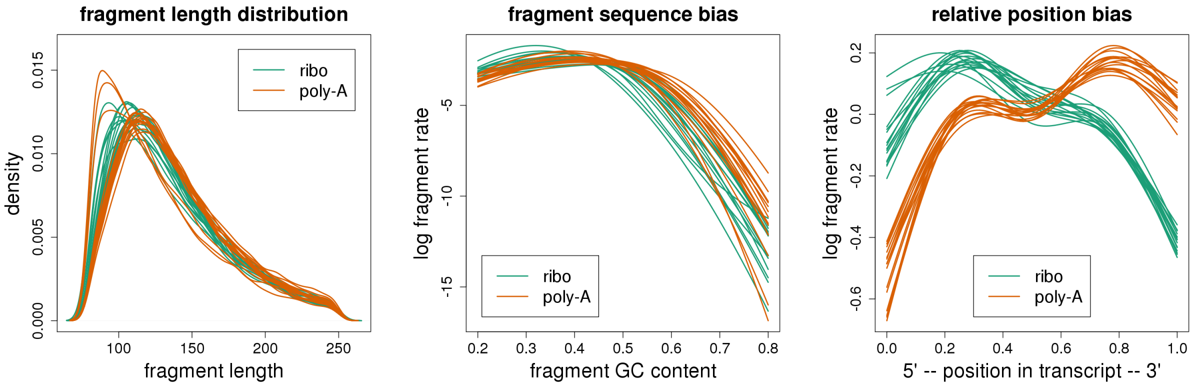
Supplementary Figure 19: Example of $\log_2$ fold error over true simulated expression for two samples and six methods. The first row shows the error rate over the true expression values for a simulated sample based on a GEUVADIS sample from sequencing center 2 (less fragment bias), and the second row for a simulated sample based on a GEUVADIS sample from center 1 (more fragment bias).



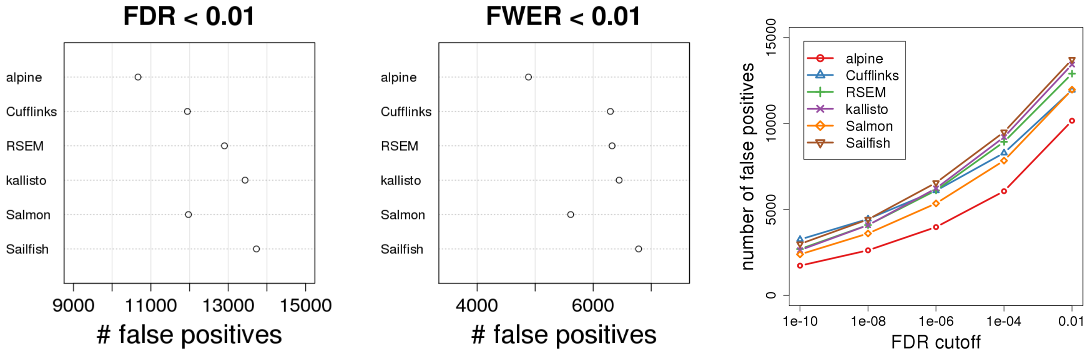
Supplementary Figure 20: Absolute value of the error in estimating percent isoform abundance for genes with two isoforms for the simulated dataset. The right panel displays the same information as the left panel, but with the y-axis scaled to 0-10% to show the middle 50% of the data for each method. The error is divided for each method into two groups separated by a black vertical line, first for the 15 samples based on sequencing center 2 (less fragment bias) and then for the 15 samples based on sequencing center 1 (more fragment bias). For each gene, a single estimate for the percent isoform abundance was calculated by averaging across 15 samples, and compared to the true percent isoform abundance. The boxplots show the distribution of errors over all simulated two-isoform genes. Most methods have increased error for the samples with more fragment bias, except *alpine* and *RSEM*, with median absolute errors across genes remaining in the range 1-2%.



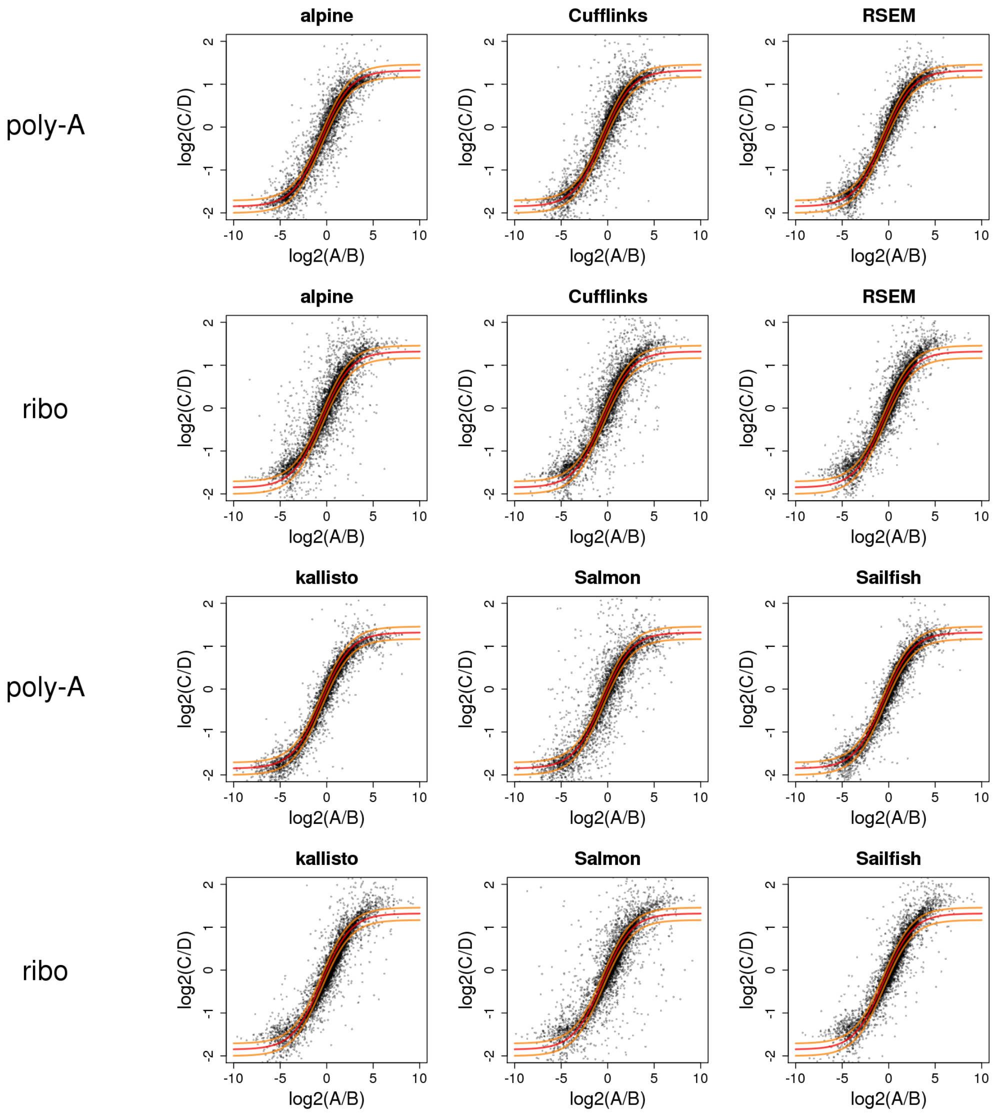
Supplementary Figure 21: Estimated bias parameters for the SEQC dataset. The top 5 panels show the fragment GC content curves for the 36 samples, combined into a single plot, split by the three sequencing locations, and lastly for library 5 which was prepared at a separate, fourth site. These panels demonstrate that the site of library preparation determines the shape of the fragment GC bias. For example, library 5 sequenced at the site "CNL" (orange curves) does not suffer from the drop-out in coverage for high GC content fragments, unlike libraries 1 and 2 which were both prepared and sequenced at that site. The bottom two panels demonstrate little difference in the fragment length distribution and positional bias across preparation or sequencing site.



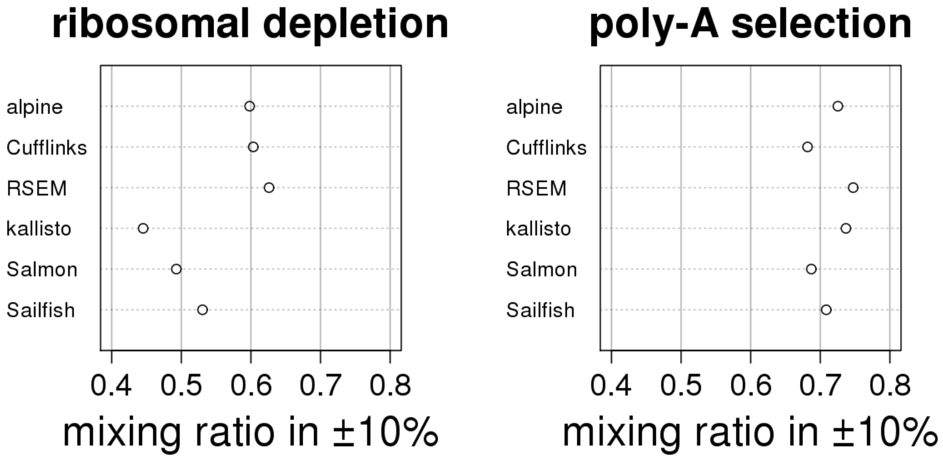
Supplementary Figure 22: Estimated bias parameters for the ABRF dataset. (Left) Fragment length distribution and (Middle) fragment GC bias curves varied little across protocol. (Right) positional bias showed a strong dependence on protocol. The expected accumulation of fragments to the 3' end of transcripts for poly-A selected libraries was observed.



Supplementary Figure 23: Number of false positives in comparing estimated transcript abundance across protocol (poly-A selection vs ribosomal RNA depletion) for samples A, B, C and D in the ABRF dataset. All methods were run with bias correction options turned on, including positional bias correction for *alpine*, *Cufflinks*, *RSEM*, and *Salmon*. (Left) Using Benjamini-Hochberg adjusted p values, controlling at 1% FDR and (Middle) using Bonferroni correction controlling at 1% FWER. (Right) Number of false positives for different methods at varying FDR cutoffs. Any differences found across protocol are categorized as false positives because it is known that the sample replicates (reference samples or defined mixtures thereof) should have identical transcript abundances. The total number of false positives at 1% FDR is extreme (out of ∼28,000 transcripts with mean *Cufflinks* FPKM > 0.1) suggesting that current computational methods for removing bias can not effectively produce protocol-invariant estimates.



Supplementary Figure 24: Mixing ratios for C/D over A/B in the ABRF dataset, split by protocol and by method for transcript abundance estimation. Transcripts in the top 25% for each method out of those with FPKM > 0.1 for both A and B samples are shown. The red line depicts the expected mixing ratio, including a correction for differences in mRNA / total RNA in samples A and B (see Methods with reference to Su et al.[11]). The oranges lines depict 10% above and below the expected mixing ratio.



Supplementary Figure 25: The fraction of transcripts with abundance estimates within 10% of the expected mixing ratio for the ABRF samples, split by protocol. For examples of the mixing ratios, see Supplementary Figure 24. Only transcripts in the top 25% of abundance estimates were used, and requiring that abundance estimates be positive for both A and B samples (FPKM > 0.1). *alpine* had relatively high recovery of the expected mixing ratio, within 5% of the top method, *RSEM* for both protocols. Over all methods, the poly-A selected samples had higher recovery of expected mixing ratio.

References

[1] Jun Li, Hui Jiang, and Wing Wong. Modeling non-uniformity in short-read rates in RNA-seq data. *Genome Biol*, 11(5):R50+, 2010.

[2] Yarden Katz, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*, 7(12): 1009–1015, 2010.

[3] Adam Roberts, Cole Trapnell, Julie Donaghey, John Rinn, and Lior Pachter. Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biol*, 12(3):R22–14, 2011.

[4] Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 2010.

[5] Rob Patro, Stephen M. Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol*, 32(5):462–464, 2014.

[6] Wei Zheng, Lisa M. Chung, and Hongyu Zhao. Bias detection and correction in RNA-sequencing data. *BMC Bioinformatics*, 12(1):290+, 2011.

[7] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*, 34(5):525–527, 2016.

[8] Rob Patro, Geet Duggal, and Carl Kingsford. Accurate, fast, and model-aware transcript expression quantification with salmon. *bioRxiv*, pages 021592+, 2015.

[9] Nicholas F. Lahens, Ibrahim H. Kavakli, Ray Zhang, Katharina Hayer, Michael B. Black, Hannah Dueck, Angel Pizarro, Junhyong Kim, Rafael Irizarry, Russell S. Thomas, Gregory R. Grant, and John B. Hogenesch. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol*, 15(6):R86+, 2014.

[10] Yarden Katz, Eric T. Wang, Jacob Silterra, Schraga Schwartz, Bang Wong, Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov, Edoardo M. Airoldi, and Christopher B. Burge. Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics*, 31(14):2400–2402, 2015.

[11] Zhenqiang Su, Paweł P. Łabaj, Sheng Li, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Shi, Charles Wang, Gary P. Schroth, Robert A. Setterquist, John F. Thompson, Wendell D. Jones, Wenzhong Xiao, Weihong Xu, Roderick V. Jensen, Reagan Kelly, Joshua Xu, Ana Conesa, Cesare Furlanello, Hanlin Gao, Huixiao Hong, Nadereh Jafari, Stan Letovsky, Yang Liao, Fei Lu, Edward J. Oakeley, Zhiyu Peng, Craig A. Praul, Javier Santoyo-Lopez, Andreas Scherer, Tieliu Shi, Gordon K. Smyth, Frank Staedtler, Peter Sykacek, Xin-Xing Tan, E. Aubrey Thompson, Jo Vandesompele, May D. Wang, Jian Wang, Russell D. Wolfinger, Jiri Zavadil, Scott S. Auerbach, Wenjun Bao, Hans Binder, Thomas Blomquist, Murray H. Brilliant, Pierre R. Bushel, Weimin Cai, Jennifer G. Catalano, Ching-Wei Chang, Tao Chen, Geng Chen, Rong Chen, Marco Chierici, Tzu-Ming Chu, Djork-Arné Clevert, Youping Deng, Adnan Derti, Viswanath Devanarayan, Zirui Dong, Joaquin Dopazo, Tingting Du, Hong Fang, Yongxiang Fang, Mario Fasold, Anita Fernandez, Matthias Fischer, Pedro Furió-Tari, James C. Fuscoe, Florian Caimet, Stan Gaj, Jorge Gandara, Huan Gao, Weigong Ge, Yoichi Gondo, Binsheng Gong, Meihua Gong, Zhuolin Gong, Bridgett Green, Chao Guo, Lei Guo, Li-Wu Guo,

James Hadfield, Jan Hellemans, Sepp Hochreiter, Meiwen Jia, Min Jian, Charles D. Johnson, Suzanne Kay, Jos Kleinjans, Samir Lababidi, Shawn Levy, Quan-Zhen Li, Li Li, Li Li, Peng Li, Yan Li, Haiqing Li, Jianying Li, Shiyong Li, Simon M. Lin, Francisco J. López, Xin Lu, Heng Luo, Xiwen Ma, Joseph Meehan, Dalila B. Megherbi, Nan Mei, Bing Mu, Baitang Ning, Akhilesh Pandey, Javier Pérez-Florido, Roger G. Perkins, Ryan Peters, John H. Phan, Mehdi Pirooznia, Feng Qian, Tao Qing, Lucille Rainbow, Philippe Rocca-Serra, Laure Sambourg, Susanna-Assunta Sansone, Scott Schwartz, Ruchir Shah, Jie Shen, Todd M. Smith, Oliver Stegle, Nancy Stralis-Pavese, Elia Stupka, Yutaka Suzuki, Lee T. Szkotnicki, Matthew Tinning, Bimeng Tu, Joost van Delft, Alicia Vela-Boza, Elisa Venturini, Stephen J. Walker, Liqing Wan, Wei Wang, Jinhui Wang, Jun Wang, Eric D. Wieben, James C. Willey, Po-Yen Wu, Jiekun Xuan, Yong Yang, Zhan Ye, Ye Yin, Ying Yu, Yate-Ching Yuan, John Zhang, Ke K. Zhang, Wenqian Zhang, Wenwei Zhang, Yanyan Zhang, Chen Zhao, Yuanting Zheng, Yiming Zhou, Paul Zumbo, Weida Tong, David P. Kreil, Christopher E. Mason, and Leming Shi. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat Biotechnol*, 32(9):903–914, 2014.